



BigML Association Discovery Cheat Sheet



Sampling

Minimum Levels for Measures

Configuration Options					
Option	Description	Default	API Name	Option	Description
Min. support	Sets a level of support between 0% and 100%. Associations below this support will be discarded.	0%	min_support	Rate	Sets the proportion of the dataset you want to consider between 0% and 100%.
Min. confidence	Sets a confidence between 0% and 100%. Associations below this confidence will be discarded.	0%	min_confidence	Range	Specifies a subset of instances from which to sample, e.g., from instance 5 to instance 1,000. The Rate you set will be computed over the Range configured.
Min. leverage	Sets a leverage between -100% and 100%. Associations below this leverage will be discarded.	0%	min_leverage	Sampling	Allows you to choose between a random sampling or a deterministic sampling. When using deterministic sampling the random number generator will always use the same seed, producing repeatable results.
Max. number (k)	Sets the maximum number of associations to be discovered. Higher numbers may take longer to calculate. You can set any value between 1 and 300.	100	max_k	Significance level	min_lift
Max. items in antecedent	Sets the maximum number of items to be considered within the antecedent itemset. You can set values between 1 and 10. The consequent itemset will always contain one item.	4	max_lis	Discretization	
Search strategy	Selects the measure to prioritize the associations discovered. Leverage is one of the measures that gives relevant results in most cases. Two other measures frequently used are confidence and lift. The strategy chosen should be coherent with your application.	Leverage	search_strategy	Option	Description
Complementary items	Takes complementary items into account. For example, if there is an item $c \circ f \rightarrow c \circ e$, its complement ($N \circ T \circ c \circ f \rightarrow c \circ e$) may also be included in some of the discovered associations. Complementary items will be represented with an exclamation point ($c \circ f \circ e \rightarrow !c \circ f \circ e$).	False	complement	Pretty	Sets segment boundaries for numeric fields, so they are easy to read. For example, instead of $segment > 20$, True pretty maximum.
Missing items	Considers missing values to be valid items, which may appear in the discovered associations.	False	missing_items	Size	Sets the number of equal segments. You can set up to 50 segments. If pretty is enabled this value acts as a maximum size.
				Trim	Sets the portion of the overall population that may be removed from either tail of the distribution. You can set a number between 0% and 10%. A trim of 1% usually gives good results.
				Type	Sets whether the field is discretized using an equal width or equal population strategy for each segment.